View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Siberian Federal University Digital Repository

УДК 519.87

ИЗВЛЕЧНИЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ АДАПТИВНЫМ МНОГОКРИТЕРИАЛЬНЫМ ГЕНЕТИЧЕСКИМ АЛГОРИТМОМ

Брестер К. Ю.
научный руководитель д-р техн. наук Семенкин Е. С.
Сибирский государственный аэрокосмический университет им. М. Ф. Решетнева

При решении задач классификации целесообразно осуществлять предобработку данных, используемых алгоритмом обучения, поскольку атрибуты могут иметь низкий уровень вариации, коррелировать друг с другом или содержать зашумленные измерения, снижающие точность классификатора.

Задача извлечения информативных признаков из базы данных (БД) может быть сформулирована в форме многокритериальной оптимизационной модели, где первый критерий f_1 представляет собой точность решения, получаемого на данной подсистеме признаков (относительная ошибка классификации), второй f_2 – число признаков, используемых для обучения классификатора:

$$\begin{cases} f_1 \to min, \\ f_2 \to min. \end{cases}$$

Оптимизация данных критериев позволяет не только улучшить качество получаемых решений, но и существенно сократить объемы используемых данных, а значит, и время, затрачиваемое на их обработку.

Предлагаемый подход (рисунок 1) основан на применении многокритериальных эволюционных технологий оптимизации для автоматического формирования системы информативных признаков. Набор имеющихся в БД признаков кодируется с помощью бинарной строки: 1 – информативный признак, 0 – неинформативный признак.

Для оценки критерия f_2 необходимо решить классификационную задачу с привлечением отобранных информативных признаков (которые соответствуют текущему варианту решения, закодированному в бинарной строке).

Существенным преимуществом является адаптация используемого метода к различным типам задач и классификаторов: для каждой задачи и модели определяется своя информативная совокупность признаков. Применение многокритериального генетического алгоритма (GA) позволяет найти компромисс между точностью распознавания и объемами используемых данных. Кроме того, для обеспечения гарантированного уровня эффективности работы, целесообразным является использование модифицированного многокритериального генетического алгоритма, основанного на идее самоадаптации. Его применение позволяет избежать настройки генетических операторов экспертом, что в свою очередь обуславливает возможность использования алгоритма для задач различного характера.

В силу свойств оптимизируемых критериев, а также высокой размерности поискового пространства для решения поставленной задачи был выбран генетический алгоритм, реализующий метод многокритериальной оптимизации SPEA (1999 г.) Приведем общую схему метода:

1. Инициализировать начальную популяцию P_0 *(t=0).*
2. Скопировать в промежуточное внешнее множество индивидов, чьи векторы решений недоминируемы относительно P_t.
3. Удалить из промежуточного внешнего множества ($\overline{P}'$) индивидов, доминируемых относительно $\overline{P}'$.

4. Если мощность $\overline{P}'$ больше заданного значения, то применить механизм кластеризации.
5. Сформировать внешнее множество из индивидов $\overline{P}'$.
6. Применить генетические операторы: селекция, скрещивание, мутация.
7. Проверить выполнение критерия останова: если выполняется – завершить работу алгоритма, иначе – перейти к п. 2.

На шаге 6 требуется настройка генетических операторов: необходимо выбрать один из вариантов скрещивания, определить вероятность мутации. В данном методе применяется турнирная селекция, причем отбор индивидов производится не только из текущей популяции, но и из внешнего множества.

В статье Дариди предложен следующий вариант адаптивной мутации:

$$p_m = \frac{1}{240} + \frac{0.11375}{2^t}, \tag{1}$$

где t – номер текущего поколения, для которого рассчитывается вероятность мутации.

Идеи коэволюционного ГА были применены для реализации адаптивного оператора скрещивания. На каждом поколении генерирование новой популяции осуществляется всеми типами скрещивания: вариантам оператора выделяются ресурсы (доля индивидов популяции, генерируемых конкретным типом скрещивания на текущем поколении) в зависимости от числа индивидов во внешнем множестве, сгенерированных при помощи данного варианта скрещивания:

$$b_i = \frac{p_i}{\left|\overline{P}\right|} \Big/ \frac{n_i}{N}, \tag{2}$$

где p_i – число индивидов во внешнем множестве, сгенерированных i-ым типом оператора скрещивания, $\left|\overline{P}\right|$ – мощность внешнего множества, n_i – число индивидов в текущей популяции, сгенерированных i-ым типом оператора, N – мощность популяции.

Для каждого варианта оператора скрещивания вычисляется «пригодность» q_i по формуле:

$$q_i = \sum_{k=0}^{T-1} \frac{T-k}{k+1} b_i, \tag{3}$$

где T – интервал адаптации, $k=0$ соответствует последнему поколению в интервале адаптации, $k=1$ – предыдущему и т.д.

Через каждые T поколений осуществляется попарное сравнение «пригодности» вариантов скрещивания с целью перераспределения ресурсов, согласно правилу:

$$s_i = \begin{cases} 0, & if \quad n_i \le social\_card \\ int\left(\dfrac{n_i - social\_card}{h_i}\right), & if \quad (n_i - h_i \cdot penalty) \le social\_card; \\ penalty, & otherwise \end{cases} \tag{4}$$

где s_i – размер ресурса, отдаваемый i-ым алгоритмом каждому победившему у него алгоритму, h_i – число поражений алгоритма в попарных сравнениях, *social_card* – минимально допустимый размер популяции, *penalty* – размер штрафа для проигравших алгоритмов. Параметр *social_card* предназначен для поддержания разнообразия вариантов оператора, *penalty* – для перераспределения ресурсов.

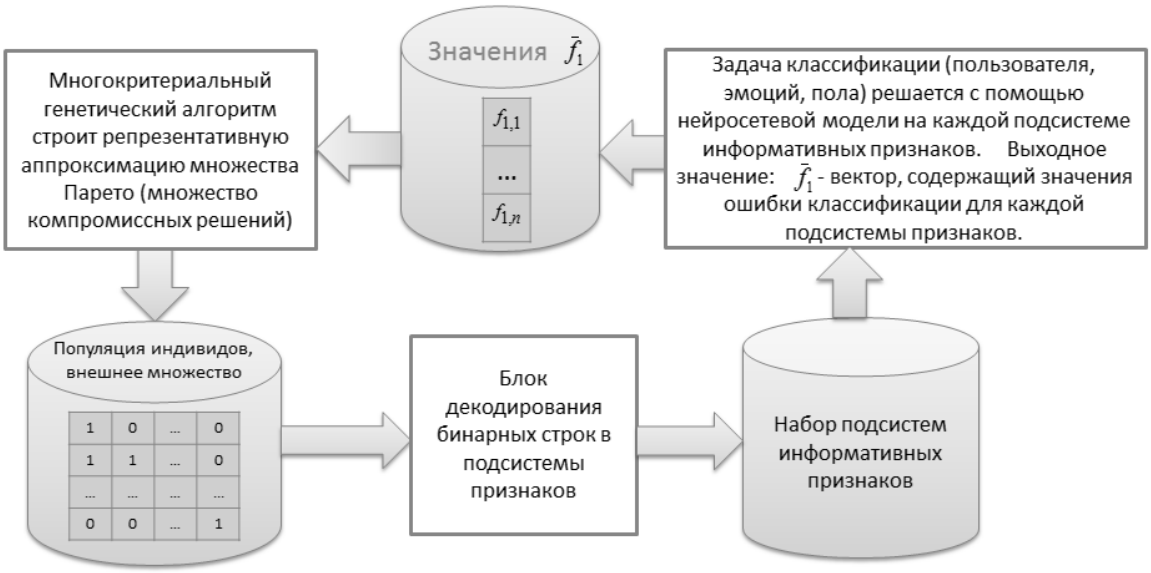

Рисунок 1 – Общая схема разрабатываемого подхода

На данном этапе в качестве классификаторов используются вероятностные нейронные сети (PNN), реализующие аппарат математической статистики (оценка плотности распределения вероятности для классов).

Разрабатываемый подход был применен к задаче распознавания эмоций говорящего по устной речи. Количество характеристик, извлекаемых из звукозаписи, достигает нескольких сотен, что существенно затрудняет работу привлекаемых алгоритмов. Поэтому важной задачей в процессе идентификации персональных характеристик человека по речи является извлечение информативной системы признаков, используемой алгоритмами распознавания.

Для тестирования предложенной алгоритмической схемы были использованы три БД: *“Berlin”*, *“SAVEE”* и *“VAM“*, содержащие характеристики голосовых записей на немецком, английском и немецком языках соответственно. Каждый звуковой файл описывался 384 признаками. Для анализа эффективности реализованного подхода для каждой задачи была получена точность классификации на полном наборе признаков, на информативной подсистеме атрибутов, извлеченной GA, и на наборе признаков, полученных с помощью метода главных компонент (PCA) (таблица 1). Полученные результаты усреднялись по 10 прогонам, выборка делилась на обучающую и тестовую в пропорции 70 на 30%.

Таблица 1 – Результаты тестирования реализованного подхода

Имя БД	PNN	PNN+GA		PNN+PCA	
	384 признака, точность (%)	Точность (%)	Количество признаков	Точность (%)	Количество признаков
Berlin	58.90	71.46	68.4	43.66	129.3
SAVEE	47.32	48.41	84.14	26.53	123.6
VAM	67.07	70.63	64.83	59.41	148.6

Анализ полученных результатов показал, что применение метода SPEA для извлечения информативных признаков позволило не только сократить число атрибутов, используемых для обучения классификатора, но и повысить точность получаемых моделей для всех БД, представленных в эксперименте. Более того разрабатываемый подход оказался эффективнее метода главных компонент, применение которого приводит к снижению точности распознавания.